Predictive Model-Checking

> print( pg.current <- poisson.gamma( 0:9, alpha.star, beta.star ) )

[1] 0.1165953406 0.2415099974 0.2587508547 0.1909752933 0.1091243547
[6] 0.0514422231 0.0208209774 0.0074357447 0.0023899565 0.0007017815

> postscript( "pg1.ps" )

> plot( 0:9, pg.current, type = 'n', xlab = 'y', ylab = 'Density' )

> for ( i in 0:9 ) {

    segments( i, 0, i, pg.current[ i + 1 ] )

}

> dev.off( )

null device

1

The omitted observed value of 0 is not too unusual in this predictive distribution.
Predictive Model-Checking

The following R code loops through the whole dataset to get the predictive $z$-scores.

```r
alpha <- beta <- 0.001
z <- rep( 0, n )
for ( i in 1:n ) {
  y.current <- y[ -i ]
  n.current <- length( y.current )
  s.current <- sum( y.current )
  alpha.star <- alpha + s.current
  beta.star <- beta + n.current
  predictive.mean.current <- alpha.star / beta.star
  predictive.SD.current <- sqrt( ( alpha.star / beta.star ) * 
                               ( 1 + 1 / beta.star ) )
  z[ i ] <- ( y[ i ] - predictive.mean.current ) / 
             predictive.SD.current
}

> z
[1] -1.43921925 -0.75757382 -0.75757382 -0.75757382 -0.75757382
[6] -0.75757382 -0.05138023 -0.05138023 -0.05138023 -0.05138023
[11]  0.68145253  0.68145253  1.44329065  3.06513271

> mean( z )
[1] 0.03133708

> sqrt( var( z ) )
[1] 1.155077
Predictive Model-Checking

> postscript( "pg2.ps" )
> qqnorm( z )
> abline( 0, 1 )

The 14 predictive \( z \)-scores have mean 0.03 (about right) and SD 1.16 (close enough to 1 when sampling variability is considered?), and the normal qqplot above shows that the only really surprising observation in the data, as far as the Poisson model was concerned, is the value of 6, which has a \( z \)-score of 3.07.

**NB** The figure above is only a crude approximation to the right qqplot, which would have to be created by simulation; even so it’s enough to suggest how the model might be improved.

I would conclude informally (a) that the Poisson is a decent model for these data, but (b) if you wanted to expand the model in a direction suggested by this diagnosis you should look for a model with extra-Poisson variation: the sample VTMR in this dataset was about 1.15.
Diffuse Priors in the LOS Case Study

In specifying a **diffuse** prior for $\lambda$ in the LOS case study, several **alternatives** to $\Gamma(\epsilon, \epsilon)$ might occur to you, including $\Gamma(1, \epsilon), \Gamma(\alpha, \beta)$ for some large $\alpha$ (like 20, to get a roughly **normal** prior) and small $\beta$ (like 1, to have a **small prior sample size**), and $U(0, C)$ for some cutoff $C$ (like 4) chosen to avoid **truncation** of the likelihood function, where $U(a, b)$ denotes the **uniform** distribution on $(a, b)$.

```r
> plot( p( lambda, 0.001, 0.001 ), lambda = 0 .. 4, v = 0 .. 0.05, color = black );
```

![Graph of uniform distribution](image1)

```r
> plot( p( lambda, 1.0, 0.001 ), lambda = 0 .. 4, color = black );
```

![Graph of Gamma distribution](image2)
Diffuse Priors (continued)

$\Gamma(1, \epsilon)$ doesn’t look promising initially as a **flat** prior, but that’s a consequence of Maple’s default choice of **vertical axis**.

> plot( p( lambda, 1.0, 0.001 ), lambda = 0 .. 4, v = 0 .. 0.05, color = black );

> plot( p( lambda, 20, 1 ), lambda = 0 .. 4, color = black );
Diffuse Priors (continued)

\[
> \text{plot( } p( \text{lambda, 20, } 1 ), \text{ lambda = 0 .. 40, color = black });
\]

\[
\Gamma(20, 1) \text{ does indeed look not far from Gaussian, and at first it may appear that it is indeed relatively flat in the region where the likelihood is appreciable } (\lambda \in (1.0, 3.3)), \text{ but we'll see below that it's actually rather more informative than we intend.}
\]

Recalling that the mean and SD of a \( \Gamma(\alpha, \beta) \) random quantity are \( \frac{\alpha}{\beta} \) and \( \sqrt{\frac{\alpha}{\beta^2}} \), respectively, and that when used as a prior with the Poisson likelihood the \( \Gamma(\alpha, \beta) \) distribution acts like a dataset with prior sample size \( \beta \), you can construct the following table:

<table>
<thead>
<tr>
<th>Prior</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>( \alpha^* )</td>
</tr>
<tr>
<td>Sample Size</td>
<td>Mean</td>
</tr>
<tr>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>1</td>
<td>0.001</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>0.001</td>
</tr>
<tr>
<td>( U(0, C) ) for ( C &gt; 4 )</td>
<td>( \frac{C}{2} )</td>
</tr>
</tbody>
</table>
Diffuse Priors (continued)

The \( \Gamma(1, \epsilon) \) prior leads to an analysis that's essentially equivalent to the integrated likelihood (fiducial) approach back on p. 72, and the \( U(0, C) \) prior for \( C > 4 \) (say) produces similar results: \( U(0, C) \) yields the \( \Gamma(s + 1, n) \) posterior truncated to the right of \( C \) (and this truncation has no effect if you choose \( C \) big enough).

You might say that the \( U(0, C) \) distribution has a prior sample size of 0 in this analysis, and its prior mean \( \frac{C}{2} \) and SD \( \frac{C}{\sqrt{12}} \) (both of which can be made arbitrarily large by letting \( C \) grow without bound) are irrelevant (an example of how intuition can change when you depart from the class of conjugate priors).

```r
> plot( { p( lambda, 29.001, 14.001 ), p( lambda, 30, 14.001 ),
       p( lambda, 49, 15 ), p( lambda, 49, 14.001 ) }, lambda = 0 .. 6,
       color = black );
```

![Graph showing posterior distributions](image)

The moral is that with only \( n = 14 \) observations, some care is needed (e.g., through pre-posterior analysis) to achieve a prior that doesn't affect the posterior very much, if that's your goal.
2.9 Continuous Outcomes

For continuous outcomes there’s an analogue of de Finetti’s Theorem that’s equally central to Bayesian model-building (e.g., Bernardo and Smith, 1994):

de Finetti’s Theorem for Continuous Outcomes.

If \( Y_1, Y_2, \ldots \) is an infinitely exchangeable sequence of real-valued random quantities with probability measure \( p \), there exists a probability measure \( Q \) over \( D \), the space of all distribution functions on \( R \), such that the joint distribution function of \( Y_1, \ldots, Y_n \) has the form

\[
p(y_1, \ldots, y_n) = \int_D \prod_{i=1}^n F(y_i) \, dQ(F), \quad (77)
\]

where \( Q(F) = \lim_{n \to \infty} p(F_n) \) and \( F_n \) is the empirical distribution function based on \( Y_1, \ldots, Y_n \).

In other words, exchangeability of real-valued observables is equivalent to the hierarchical model

\[
F \sim p(F) \quad \text{(prior)}
\]

\[
(Y_1, \ldots, Y_n \mid F) \overset{\text{iid}}{\sim} F \quad \text{(likelihood)} \quad (78)
\]

for some prior distribution \( p \) on the set \( D \) of all possible distribution functions.

This prior makes the continuous form of de Finetti’s Theorem considerably harder to apply: to take the elicitation task seriously is to try to specify a probability distribution on a function space (\( F \) is in effect an infinite-dimensional parameter).

(NB This task is not unique to Bayesians—it’s equivalent to asking “Where does the likelihood come from?” in frequentist analyses of observational data.)
Continuous Outcomes (continued)

What people often do in practice is to appeal to considerations that narrow down the field, such as an a priori judgment that the $Y_i$ ought to be symmetrically distributed about a measure of center $\mu$, and then try to use a fairly rich parametric family satisfying (e.g.) the symmetry restriction as a substitute for all of $D$.

Strictly speaking you’re not supposed to look at the $Y_i$ while specifying your prior on $D$—this can lead to a failure to fully assess and propagate model uncertainty—but not doing so can permit the data to surprise you in ways that would make you want to go back and revise your prior (an example of Cromwell’s Rule in action).

As mentioned earlier, I’ll suggest two potential ways out of this dilemma, based on out-of-sample predictive validation (the model-checking in the LOS data above was an example of this) and Bayesian nonparametrics.

Case Study: Measurement of physical constants. What used to be called the National Bureau of Standards (NBS) in Washington, DC, conducts extremely high precision measurement of physical constants, such as the actual weight of so-called check-weights that are supposed to serve as reference standards (like the official kg).

In 1962–63, for example, $n = 100$ weighings (listed below) of a block of metal called NB10, which was supposed to weigh exactly 10g, were made under conditions as close to IID as possible (Freedman et al., 1998).

<table>
<thead>
<tr>
<th>Value</th>
<th>375</th>
<th>392</th>
<th>393</th>
<th>397</th>
<th>398</th>
<th>399</th>
<th>400</th>
<th>401</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>7</td>
<td>4</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value</th>
<th>402</th>
<th>403</th>
<th>404</th>
<th>405</th>
<th>406</th>
<th>407</th>
<th>408</th>
<th>409</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>8</td>
<td>6</td>
<td>9</td>
<td>5</td>
<td>12</td>
<td>8</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value</th>
<th>410</th>
<th>411</th>
<th>412</th>
<th>413</th>
<th>415</th>
<th>418</th>
<th>423</th>
<th>437</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

89
NB10 Modeling

Q: (a) How much does NB10 really weigh? (b) How certain are you given the data that the true weight of NB10 is less than (say) 405.25? And (c) How accurately can you predict the 101st measurement?

The graph below is a normal qqplot of the 100 measurements \( y = (y_1, \ldots, y_n) \), which have a mean of \( \bar{y} = 404.6 \) (the units are micrograms below 10g) and an SD of \( s = 6.5 \).

![Normal QQ Plot]

Evidently it’s plausible in answering these questions to assume symmetry of the “underlying distribution” \( F \) in de Finetti’s Theorem.

One standard choice, for instance, is the Gaussian:

\[
(\mu, \sigma^2) \sim p(\mu, \sigma^2) \quad \text{IID} \quad (Y_i|\mu, \sigma^2) \sim N(\mu, \sigma^2). \tag{79}
\]

Here \( N(\mu, \sigma^2) \) is the familiar normal density

\[
p(y_i|\mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{y_i - \mu}{\sigma} \right)^2 \right]. \tag{80}
\]
Gaussian Modeling

Even though you can see from the previous graph that (79) is not a good model for the NB10 data, I'm going to fit it to the data for practice in working with the normal distribution from a Bayesian point of view (later we'll improve upon the Gaussian).

(79) is more complicated than the models in the AMI and LOS case studies because the parameter $\theta$ here is a vector:

$$\theta = (\mu, \sigma^2).$$

To warm up for this new complexity let's first consider a cut-down version of the model in which we pretend that $\sigma$ is known to be $\sigma_0 = 6.5$ (the sample SD).

This simpler model is then

$$\left\{ \begin{array}{c} \mu \\ (Y_i|\mu) \end{array} \right\} \sim p(\mu) \quad \text{IID} \quad N(\mu, \sigma_0^2).$$

The likelihood function in this model is

$$l(\mu|y) = \prod_{i=1}^{n} \frac{1}{\sigma_0 \sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_0^2}(y_i - \mu)^2\right]$$

$$= c \exp\left[-\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (y_i - \mu)^2\right]$$

$$= c \exp\left[-\frac{1}{2\sigma_0^2} \left( \sum_{i=1}^{n} y_i^2 - 2\mu \sum_{i=1}^{n} y_i + n\mu^2 \right)\right]$$

$$= c \exp\left[-\frac{1}{2 \left(\frac{\sigma_0^2}{n}\right)} (\mu - \bar{y})^2 \right].$$

Thus the likelihood function, when thought of as a density for $\mu$, is a normal distribution with mean $\bar{y}$ and SD $\frac{\sigma_0}{\sqrt{n}}$. 

91
Gaussian Modeling (continued)

Notice that this SD is the same as the frequentist **standard error** for \( \bar{Y} \) based on an IID sample of size \( n \) from the \( N(\mu, \sigma^2_0) \) distribution.

(82) also shows that the sample mean \( \bar{y} \) is a **sufficient statistic** for \( \mu \) in model (81).

In finding the conjugate prior for \( \mu \) it would be nice if the **product of two normal distributions is another normal distribution**, because that would demonstrate that the conjugate prior is normal.

Suppose therefore, to see where it leads, that the **prior for** \( \mu \) is (say) \( p(\mu) = N(\mu_0, \sigma^2_\mu) \).

Then **Bayes’ Theorem** would give

\[
p(\mu|y) = c p(\mu) l(\mu|y) \tag{83}
\]

\[
= c \exp \left[ -\frac{1}{2\sigma^2_\mu} (\mu - \mu_0)^2 \right] \exp \left[ -\frac{n}{2\sigma^2_0} (\mu - \bar{y})^2 \right]
\]

\[
= c \exp \left\{ -\frac{1}{2} \left[ \frac{(\mu - \mu_0)^2}{\sigma^2_\mu} + \frac{n(\mu - \bar{y})^2}{\sigma^2_0} \right] \right\},
\]

and we want this to **be of the form**

\[
p(\mu|y) = c \exp \left\{ -\frac{1}{2} \left[ A(\mu - B)^2 + C \right] \right\}
\]

\[
= c \exp \left\{ -\frac{1}{2} \left[ A\mu^2 - 2AB\mu + (AB^2 + C) \right] \right\} \tag{84}
\]

for some \( B, C, \) and \( A > 0. \)

Maple can help **see if this works**:

\[
> \text{collect}( (\mu - \mu_0)^2 / \text{sigmu}^2 + n * (\mu - \bar{y})^2 / \text{sigma0}^2, \mu, \mu_0, \text{mu0}, \text{n ybar} );
\]

\[
\frac{1}{n} \left[ \frac{1}{\text{mu0}} - \frac{1}{\text{sigmu}} \right] \frac{1}{\text{mu0}^2} + \frac{n}{\text{sigma0}} \frac{1}{\text{ybar}} \frac{1}{\text{sigma0}^2} + \frac{2}{\text{mu0}} \frac{n}{\text{ybar}} \frac{2}{\text{sigma0}^2} + \frac{2}{\text{sigmu}} \frac{\text{mu0}}{\text{sigma0}^2} + \frac{2}{\text{sigmu}} \frac{n}{\text{ybar}} \frac{2}{\text{sigma0}^2}
\]
Gaussian Modeling

Matching coefficients for $A$ and $B$ (we don’t really care about $C$) gives

$$A = \frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_0^2} \quad \text{and} \quad B = \frac{\mu_0}{\sigma_\mu^2} + \frac{n\bar{y}}{\sigma_o^2}. \quad (85)$$

Since $A > 0$ this demonstrates two things: (1) the conjugate prior for $\mu$ in model (81) is normal, and (2) the conjugate updating rule (when $\sigma_0$ is assumed known) is

$$\left\{ \begin{array}{l}
\mu \sim N(\mu_0, \sigma_\mu^2) \\
(Y_i|\mu) \overset{\text{IID}}{\sim} N(\mu, \sigma_0^2), \\
 i = 1, \ldots, n
\end{array} \right\} \rightarrow (\mu|y) = (\mu|\bar{y}) = N(\mu_*, \sigma_*) , \quad (86)$$

where the posterior mean and variance are given by

$$\mu_* = B = \left( \frac{1}{\sigma_\mu^2} \right) \mu_0 + \left( \frac{n}{\sigma_0^2} \right) \bar{y} \quad \text{and} \quad \sigma_*^2 = A^{-1} = \frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_0^2}. \quad (87)$$

It becomes useful in understanding the meaning of these expressions to define the precision of a distribution, which is just the reciprocal of its variance: whereas the variance and SD scales measure uncertainty, the precision scale quantifies information about an unknown.

With this convention (87) has a series of intuitive interpretations, as follows:

- The prior, considered as an information source, is Gaussian with mean $\mu_0$, variance $\sigma_\mu^2$, and precision $\frac{1}{\sigma_\mu^2}$, and when viewed as a data set consists of $n_0$ (to be determined below) observations;

- The likelihood, considered as an information source, is Gaussian with mean $\bar{y}$, variance $\frac{\sigma_0^2}{n}$, and precision $\frac{n}{\sigma_0^2}$, and when viewed as a data set consists of $n$ observations;
Gaussian Modeling (continued)

- The **posterior**, considered as an **information source**, is Gaussian, and the posterior mean is a **weighted average** of the prior mean and data mean, with weights given by the **prior** and **data precisions**;

- The **posterior precision** (the reciprocal of the posterior variance) is just the **sum** of the prior and data precisions (this is why people invented the idea of precision—on this scale **knowledge** about $\mu$ in model (81) is **additive**); and

- **Rewriting $\mu_*$ as**

$$
\mu_* = \frac{\left( \frac{1}{\sigma_*^2} \right) \mu_0 + \left( \frac{n}{\sigma_0^2} \right) \bar{y}}{\frac{1}{\sigma_*^2} + \frac{n}{\sigma_0^2}} = \frac{\left( \frac{\sigma_0^2}{\sigma_*^2} \right) \mu_0 + n \bar{y}}{\frac{\sigma_0^2}{\sigma_*^2} + n},
$$

(88)

you can see that the **prior sample size** is

$$
n_0 = \frac{\sigma_0^2}{\sigma_*^2} = \frac{1}{\left( \frac{\sigma_*}{\sigma_0} \right)^2},
$$

(89)

which makes sense: the **bigger** $\sigma_*$ is in relation to $\sigma_0$, the **less prior information** is being incorporated in the conjugate updating (86).

**Bayesian inference with multivariate $\theta$.** Returning now to (79) with $\sigma^2$ unknown, (as mentioned above) this model has a ($p = 2$)-dimensional **parameter vector** $\theta = (\mu, \sigma^2)$.

When $p > 1$ you can still use Bayes’ Theorem directly to obtain the **joint posterior distribution**, 

$$
p(\theta | y) = p(\mu, \sigma^2 | y) = c p(\theta) l(\theta | y) \\
= c p(\mu, \sigma^2) l(\mu, \sigma^2 | y),
$$

(90)
Multivariate Unknown $\theta$

where $y = (y_1, \ldots, y_n)$, although making this calculation directly requires a $p$-dimensional integration to evaluate the normalizing constant $c$; for example, in this case

$$
c = [p(y)]^{-1} = \left( \int \int p(\mu, \sigma^2, y) \, d\mu \, d\sigma^2 \right)^{-1}
$$

$$
= \left( \int \int p(\mu, \sigma^2) l(\mu, \sigma^2|y) \, d\mu \, d\sigma^2 \right)^{-1}.
$$

(91)

Usually, however, you’ll be more interested in the marginal posterior distributions, in this case $p(\mu|y)$ and $p(\sigma^2|y)$.

Obtaining these requires $p$ integrations, each of dimension $(p - 1)$, a process that people refer to as marginalization or integrating out the nuisance parameters—for example,

$$
p(\mu|y) = \int_0^\infty p(\mu, \sigma^2|y) \, d\sigma^2.
$$

(92)

**Predictive** distributions also involve a $p$-dimensional integration: for example, with $y = (y_1, \ldots, y_n),$

$$
p(y_{n+1}|y) = \int \int p(y_{n+1}, \mu, \sigma^2|y) \, d\mu \, d\sigma^2
$$

(93)

$$
= \int \int p(y_{n+1}|\mu, \sigma^2) \, p(\mu, \sigma^2|y) \, d\mu \, d\sigma^2.
$$

And, finally, if you’re interested in a function of the parameters, you have some more hard integrations ahead of you.

For instance, suppose you wanted the posterior distribution for the *coefficient of variation* $\lambda = g_1(\mu, \sigma^2) = \frac{\sqrt{\sigma^2}}{\mu}$ in model (79).
Multivariate Unknown $\theta$

Then one fairly direct way to get this posterior (e.g., Bernardo and Smith, 1994) is to (a) introduce a second function of the parameters, say $\eta = g_2(\mu, \sigma^2)$, such that the mapping $f = (g_1, g_2)$ from $(\mu, \sigma^2)$ to $(\lambda, \eta)$ is invertible; (b) compute the joint posterior for $(\lambda, \eta)$ through the usual change-of-variables formula

$$p(\lambda, \eta | y) = p_{\mu, \sigma^2}[f^{-1}(\lambda, \eta) | y] \cdot |J_{f^{-1}}(\lambda, \eta)|,$$  \hspace{1cm} (94)

where $p_{\mu, \sigma^2}(\cdot, \cdot | y)$ is the joint posterior for $\mu$ and $\sigma^2$ and $|J_{f^{-1}}|$ is the determinant of the Jacobian of the inverse transformation; and (c) marginalize in $\lambda$ by integrating out $\eta$ in $p(\lambda, \eta | y)$, in a manner analogous to (92).

Here, for instance, $\eta = g_2(\mu, \sigma^2) = \mu$ would create an invertible $f$, with inverse defined by $(\mu = \eta, \sigma^2 = \lambda^2 \eta^2)$; the Jacobian determinant comes out $2\lambda \eta^2$ and (94) becomes

$$p(\lambda, \eta | y) = 2\lambda \eta^2 \cdot p_{\mu, \sigma^2}(\eta, \lambda^2 \eta^2 | y).$$

This process involves two integrations, one (of dimension $p$) to get the normalizing constant that defines (94) and one (of dimension $(p - 1)$) to get rid of $\eta$.

You can see that when $p$ is a lot bigger than 2 all these integrals may create severe computational problems—this has been the big stumbling block for applied Bayesian work for a long time.

More than 200 years ago Laplace (1774)—perhaps the second applied Bayesian in history (after Bayes himself)—developed, as one avenue of solution to this problem, what people now call Laplace approximations to high-dimensional integrals of the type arising in Bayesian calculations (see, e.g., Tierney and Kadane, 1986).

Starting in the next case study after this one, we’ll use another, computationally intensive, simulation-based approach: Markov chain Monte Carlo (MCMC).
**Gaussian Modeling**

**Back to model (79).** The conjugate prior for $\theta = (\mu, \sigma^2)$ in this model (e.g., Gelman et al., 2003) turns out to be most simply described **hierarchically**:

\[
\begin{align*}
\sigma^2 & \sim \text{SI-}\chi^2(\nu_0, \sigma_0^2) \\
(\mu|\sigma^2) & \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right). \tag{95}
\end{align*}
\]

Here saying that $\sigma^2 \sim \text{SI-}\chi^2(\nu_0, \sigma_0^2)$, where SI stands for **scaled inverse**, amounts to saying that the precision $\tau = \frac{1}{\sigma^2}$ follows a **scaled** $\chi^2$ distribution with parameters $\nu_0$ and $\sigma_0^2$.

The scaling is chosen so that $\sigma_0^2$ can be interpreted as a **prior estimate** of $\sigma^2$, with $\nu_0$ the **prior sample size** of this estimate (i.e., **think of a prior data set with** $\nu_0$ **observations and sample SD** $\sigma_0$).

Since $\chi^2$ is a special case of the Gamma distribution, SI-$\chi^2$ must be a special case of the **inverse Gamma** family—its **density** (see Gelman et al., 2003, Appendix A) is

\[
\begin{align*}
\sigma^2 & \sim \text{SI-}\chi^2(\nu_0, \sigma_0^2) \iff \\
p(\sigma^2) & = \frac{(\frac{1}{2}\nu_0)^{\frac{1}{2}\nu_0}}{\Gamma\left(\frac{1}{2}\nu_0\right)} \left(\sigma_0^2\right)^{\frac{1}{2}\nu_0} (\sigma^2)^{-\left(1+\frac{1}{2}\nu_0\right)} \exp\left(-\frac{\nu_0 \sigma_0^2}{2\sigma^2}\right). \tag{96}
\end{align*}
\]

As may be verified with **Maple**, this distribution has **mean** (provided that $\nu_0 > 2$) and **variance** (provided that $\nu_0 > 4$) given by

\[
\begin{align*}
E(\sigma^2) & = \frac{\nu_0}{\nu_0 - 2} \sigma_0^2 \quad \text{and} \quad V(\sigma^2) = \frac{2\nu_0^2}{(\nu_0 - 2)^2(\nu_0 - 4)} \sigma_0^4. \tag{97}
\end{align*}
\]
Gaussian Modeling (continued)

The parameters $\mu_0$ and $\kappa_0$ in the second level of the prior model (95), $(\mu|\sigma^2) \sim N(\mu_0, \kappa_0 \sigma^2)$, have **simple parallel interpretations** to those of $\sigma_0^2$ and $\nu_0$: $\mu_0$ is the prior estimate of $\mu$, and $\kappa_0$ is the **prior effective sample size** of this estimate.

The **likelihood function** in model (79), with both $\mu$ and $\sigma^2$ unknown, is

$$
l(\mu, \sigma^2|y) = c \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{1}{2\sigma^2} (y_i - \mu)^2 \right]
$$

$$
= c (\sigma^2)^{-\frac{1}{2}n} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right] 
$$

$$
= c (\sigma^2)^{-\frac{1}{2}n} \exp \left[ -\frac{1}{2\sigma^2} \left( \sum_{i=1}^{n} y_i^2 - 2\mu \sum_{i=1}^{n} y_i + n\mu^2 \right) \right].
$$

The **expression in brackets** in the last line of (98) is

$$
\left[ \cdot \right] = -\frac{1}{2\sigma^2} \left( \sum_{i=1}^{n} y_i^2 + n(\mu - \bar{y})^2 - n\bar{y}^2 \right) 
$$

$$
= -\frac{1}{2\sigma^2} [n(\mu - \bar{y})^2 + (n - 1)s^2],
$$

where $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the **sample variance**. Thus

$$
l(\mu, \sigma^2|y) = c (\sigma^2)^{-\frac{1}{2}n} \exp \left\{ -\frac{1}{2\sigma^2} [n(\mu - \bar{y})^2 + (n - 1)s^2] \right\},
$$

and it's clear that the vector $(\bar{y}, s^2)$ is **sufficient** for $\theta = (\mu, \sigma^2)$ in this model, i.e., $l(\mu, \sigma^2|y) = l(\mu, \sigma^2|\bar{y}, s^2)$.  

98
Gaussian Analysis

Maple can be used to make \textbf{3D} and \textbf{contour plots} of this likelihood function with the NB10 data:

\begin{verbatim}
> l := ( mu, sigma2, ybar, s2, n ) -> sigma2^(- n / 2) * exp(- ( n * ( mu - ybar )^2 + ( n - 1 ) * s2 ) / ( 2 * sigma2 ) );

l := (mu, sigma2, ybar, s2, n) ->

\[
\frac{2}{\text{sigma2}} \exp\left(- \frac{1}{2} \frac{n (\mu - ybar) + (n - 1) s2}{\text{sigma2}}\right)
\]

> plotsetup( x11 );

> plot3d( l( mu, sigma2, 404.6, 42.25, 100 ), mu = 402.6 .. 406.6, sigma2 = 25 .. 70 );
\end{verbatim}

You can use the mouse to \textbf{rotate} \textbf{3D} plots and get \textbf{other useful views} of them:
The projection or shadow plot of $\mu$ looks a lot like a normal (or maybe a $t$) distribution.

And the shadow plot of $\sigma^2$ looks a lot like a Gamma (or maybe an inverse Gamma) distribution.